On the Concept of Correct Hits in Spoken Term Detection*

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences and University of Szeged
H-6720 Szeged, Tisza Lajos krt. 103., Hungary
ggabor@inf.u-szeged.hu

Abstract. In most Information Retrieval (IR) tasks the aim is to find human-comprehensible items of information in large archives. One such task is the spoken term detection (STD) one, where we look for user-entered keywords in a large audio database. To evaluate the performance of a spoken term detection system we have to know the real occurrences of the keywords entered. Although there are standard automatic ways to obtain these locations, it is not obvious how these match user expectations. In our study, we asked a number of subjects to locate these relevant occurrences, and we compared the performance of our spoken term detection system using their responses. In addition, we investigated the nature of their answers, seeking to find a way to determine a commonly accepted list of relevant occurrences.

KeyWords: spoken term detection, information retrieval, artificial intelligence, speech processing, keyword spotting

Spoken term detection [19] is a relatively new area, which is closely related to speech recognition. Both seek to precisely match the relation between audio speech recordings and their transcripts; but while speech recognition seeks to produce the correct transcript of speech utterances [16], spoken term detection attempts to locate those parts of the utterance where the user-entered keyword or keywords occur.

One critical part of the latter concept is that of identifying the relevant occurrences of the keywords. At first glance, this question could be answered quite easily, provided we have the correct, time-aligned textual representation (*transcription*) of the utterances: a standard solution is to consider an occurrence relevant if it is present at the given position as a whole word [15]. However, this approach completely ignores compound words, which could also be considered relevant occurrences. A further problem arises in the *agglutinative* languages [7, 4]: these construct new word forms by adding affix morphemes to the end of

* This publication is supported by the European Union and co-funded by the European Social Fund. Project title: Telemedicine-focused research activities in the fields of mathematics, informatics and medical sciences. Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073.

the word stem. (E.g. in Hungarian the expression "in my house" takes the form *ház-am-ban*.) In these cases, inflected forms of keywords should also be accepted. (This is even present in English to a certain extent, e.g. the plural form of nouns.)

The best solution for this task would be to ask the user which occurrences he thinks are relevant. The problem with this approach is that usually the archives are huge, hence hand-labeling them is quite expensive. Furthermore, the expectations could vary from user to user, but for practical reasons we would need an "objective" list of the relevant occurrences. It is also not clear whether, by using user responses, a broad consensus could be reached; i.e. whether it is possible to create an occurrence list that is acceptable to most people.

In this study we examined these expectations, and we also sought to measure the effect of these on STD accuracy. (Although we think that the topic of this paper is not limited to spoken term detection, but it also covers several IR topics like text document retrieval [3] and document categorization [20] as well.) For this reason, we created a form containing ambiguous occurrences and asked people about their opinions of relevance. The results were compared with each other, and with our standard, automatic occurrence-detection method.

Although our experiments were performed on a set of Hungarian recordings, we think that our findings might be of interest to researchers working with other languages as well, especially as recently languages other than English have been receiving more attention (e.g. [14, 17, 22, 13]).

1 The Spoken Term Detection Task

In the spoken term detection task we would like to find the user-entered expressions (called *terms* or *keywords*) in an audio database (the set of *recordings*). An STD method returns a list of *hits*, each consisting of the position of occurrence (a speech signal index, starting and ending times), the term found, and a probability value that can be used to rank the hits. In contrast to other similar tasks, in STD the order of the hits does not matter; the probability value is primarily used to further filter the hit list, keeping just the more probable elements.

As a user expects a quick response for his input, we have to scan hours of recordings in just a few seconds (or less); to achieve this, the task is usually separated into two distinct parts. In the first one, steps requiring intensive computation are performed without knowing the actual search term, resulting in some intermediate representation. Then, when the user enters the keyword(s): some kind of (quick) search is performed in this representation. There exist a number of such intermediate representations, from which we used the one where we stored only the most probable phoneme sequence for a recording [15, 6].

In this paper we will concentrate on the concept of *relevant occurrence*; hence spoken term detection is only of interest to us here because it can provide us with accuracy scores that can be compared with each other when using different strategies for detecting these occurrences. Therefore, in a quite unusual way, we will use the *same* STD system configuration, with exactly the same parameters; what we will vary is the occurrences of search terms we expect it to find.

1.1 The Evaluation Metrics

A spoken term detection system returns a list of hits for a query. Given the correct list of occurrences, we should rate the performance of the system to be able to compare different systems and configurations. Since STD is an information retrieval task, it is straightforward to apply standard IR metrics of precision and recall:

$$Precision = \frac{N_C}{N_C + N_{FA}} \tag{1}$$

and

$$Recall = \frac{N_C}{N_{Total}}, \tag{2}$$

where N_C is the number of correct hits returned, N_{FA} is the number of false alarms, and N_{Total} is the total number of real occurrences [1]. Intuitively, precision measures how much of the hit list returned contains correct hits, while recall measures the fraction of the real occurrences that were found. A perfect system has both a precision and a recall score of 1 (or 100%). Clearly, there is a trade-off between these two values: high precision can easily lead to a low recall score if we only include very probable hits in our list, while it is easy to achieve high recall rates and get poor precision scores by returning a hit list full of "rubbish". Hence it would be better to summarize the performance of a system using just one score. In IR tasks usually the F-measure (or F_1) is used for this, which is the harmonic mean of precision and recall, defined as

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{3}$$

This formula, however, weights precision and recall equally, which might differ from our preferences. We could also use different weights for the two measures, but their relative importance is not really clear. Another requirement in STD might be to normalize the scores based on the total length of the recordings. This is why in the area of spoken term detection usually some other – although similar – measures are used.

Figure-of-Merit (FOM) The evaluation metric commonly applied earlier is the Figure-of-Merit (FOM). It can be calculated simply as the mean of the recall scores when we allow only $1, 2, \ldots 10$ false alarms per hour per keyword. In general, this metric is a quite permissive one: it is possible to achieve relatively high scores quite easily, since 10 false alarms per hour clearly exceeds the limits of actual applicability. It weights keywords relative to their frequency of occurrence in the archive of recordings, hence if we want to maximize this score, it may be worth optimizing it on more frequent keywords instead of rarer ones. However, this behaviour is clearly contrary to user expectations. Another interesting property is that the STD system does not have to filter the hits returned, but the FOM metric determines the actual probability thresholds depending on the number of false alarms permitted.

Actual Term-Weighted Value (ATWV) Another, more strict measure was defined by the National Institute of Standards and Technology (NIST) in its 2006 evaluation of Spoken Term Detection [12]. Unlike FOM, it uses all the hits supplied by the STD method, and is defined as

$$ATWV = 1 - \frac{1}{T}\sum_{t=1}^{T}\big(P_{Miss}(t) + \beta P_{FA}(t)\big), \tag{4}$$

where T is the number of terms, $P_{Miss}(t)$ is the probability value of missing the term t and $P_{FA}(t)$ is the probability value of a false alarm. These probability values are defined as

$$P_{Miss}(t) = 1 - \frac{N_C(t)}{N_{Total}(t)} \tag{5}$$

and

$$P_{FA}(t) = \frac{N_{FA}(t)}{T_{speech} - N_{Total}(t)}, \tag{6}$$

where T_{speech} is the duration of the test speech in seconds. (This formula uses the somewhat arbitrary assumption that every term can occur once in every second.) Usually the penalty factor for false alarms (β) is set to 1000. A system achieving perfect detection (i.e. having a precision and a recall of 1.0) has an ATWV score of 1.0; a system returning no hits has a score of 0.0; while a system which finds all occurrences, but produces 3.6 false alarms for each term and speech hour also has a score of 0.0 (assuming that T_{speech} is significantly larger than N_{Total}) [15].

ATWV differs from FOM in a number of ways. First, it weights all keywords equally, regardless of the frequency of actual occurrences. Second, it punishes missed occurrences and false alarms much more than FOM does, so it is a very strict metric indeed. Third, whereas FOM performs a filtering of the hits returned, ATWV uses *all* of them, hence to achieve a high ATWV score an STD system has to filter the hit lists itself by setting up a minimal probability threshold $P_{\min}$. This is usually done in two steps: first the actual $P_{\min}$ value is determined on a *development set* of recordings as the threshold value belonging to the optimal ATWV score. Then, to measure actual performance, ATWV is calculated on another set of recordings (the *test set*) using the already determined value for $P_{\min}$. In this study, we also performed these two steps.

2 The Concept of "Correct Hit"

Having defined the evaluation metrics, we are now able to calculate the accuracy scores of an STD system when we have the number of correct hits, false alarms and missed occurrences. For this, we supposedly know the hits returned, and their ordering; however, we still have to define the technique to get the list of real occurrences, and the way of matching returned hits and the relevant occurrences.

2.1 Matching Hits and Occurrences

In the literature this topic has been discussed quite extensively. Of course, a hit and an occurrence can be matched only if the keywords are the same, and they occur in the same recording. As regards the match of time-alignment, there are a number of possibilities. A valid option would be to expect both the starting and ending times to lie below a threshold. [12] expects the time span of the hit to be in at most 0.5 seconds from the centre of the real occurrence. [21] demanded that the time spans of the hit and of the occurrence intersect. We chose the latter method, partly because of the agglutinative nature of the Hungarian language, which makes the task of determining the exact keyword starting and ending times quite hard.

2.2 Determining the "Real" Occurrences

When we search for the method of choosing the "relevant occurrences" in the literature, we usually find no mention of it. Hence we chose to assume that a keyword only occured if it was present in the textual transcription as a whole word by itself. This approach, however, is hardly applicable when we work with recordings different from English (which was also the case for us). In morphologically rich languages such as Hungarian, nouns (which are typical candidates for keywords) can have hundreds of different forms owing to grammatical number, possession marking and grammatical cases, all of these forms being ones that should also be treated as "real" occurrences.

Our standard automatic method is a simple variation of this default approach. In it we treat a given position as an occurrence of the given keyword if the word at this position contains the keyword. (This concept can be extended to keywords consisting of several words in a straightforward way.) Because in Hungarian a noun ending with a vowel may change its form when getting some inflections (like the noun *"Amerika"* (*America*) changing to the form *"Amerikában"* (*in America*)), we also considered the occurrence a real one if the given keyword appears in the form having its last vowel substituted by its long counterpart, as long as the last vowel is also the last phoneme of the keyword.

It is of course known that this technique is not perfect: for short keywords in particular it is likely that they will appear inside other words having a completely different meaning, which should be categorized as false alarms.

2.3 Relying on Human Expectations

The other choice is to employ the concept that a relevant occurrence is where the actual users think that the current occurrence is indeed relevant. This approach sounds quite reasonable, but it requires valuable human interaction, so it could be quite labour-intensive when we have to annotate a big archive manually. For smaller archives, however, it can be carried out relatively cheaply; and since the aim of this study was to check the difference between the automatic and human concept of a real occurrence, we performed this manual task by asking our subjects about their opinions of potential occurrences.

Strategy	Dev	Test
Automatic	381	709
Subject #1	365	690
Subject #2	368	689
Subject #3	396	732
Subject #4	366	699
Subject #5	367	697
Subjects (majority voting)	367	697
Clean occurrences	334	651

Table 1. The number of relevant occurrences using different strategies for determining correct hits for the development (*Dev*) and test (*Test*) sets.

Creating the Form to Fill In To make subjects list the occurrences which they thought were relevant, we created a form using the textual transcript of the recordings, which each subject had to fill in. For each keyword we located the similar letter-sequences in the transcripts of the recordings using the edit distance [9]: we allowed character insertions, deletions and substitutions, and listed the parts of the recordings where we could reproduce the given keyword with at most N operations, where N was 30% of the length of the keyword. (That is, for a search term consisting of 10 characters, we allowed only 3 operations.)

Because this list was still quite long, we shortened it with a simple trick: we did not list those occurrences which could be produced without any operations, and were located at the beginning of a word. Instead, we assumed that these were the occurrences of the actual keyword in inflected form, thus treating them as relevant occurrences. (The set of these occurrences was also used in the experiments section, referred to as the list of *clean occurrences*.) Of course this was not so in a number of cases (like certain compound words), but this technique was quite close to our objective, and it effectively reduced the number of items in the form.

Evaluating Subject Responses Table 1 shows the number of relevant occurrences found when using the automatic occurrence detector method (see line "Automatic"), and for the responses of the subjects (see lines "Subject #N"). The form contained 111 (development set) and 242 (test set) occurrences that were used to decide on their relevance; from these, the test subjects marked between 31 and 62, and between 38 and 81 occurrences as relevant ones, development sets and test sets, respectively. The results indicate that most occurrences were judged in quite a similar way by our subjects (with the exception of Subject #3). Besides comparing the responses of the subjects with the results of our standard automatic occurrence checker, we also wanted to know whether a consensus could be reached between the answers of the subjects. For this reason we used majority voting: we considered an occurrence relevant if at least half of the subjects (now at least three of them) considered it relevant.

3 Experiments and Results

Having defined the task, introduced the method of obtaining subject responses, and selected the evaluation metrics, we will now turn to the testing part. We will describe the STD framework used, present and analyze the results, concentrating on the various kinds of discrepancies among the individual subjects, and between each subject and the automatic occurrence detector method used.

3.1 The STD Framework

Testing was performed using the spoken term detection system presented in [6]. It uses phoneme sequences as an intermediate representation, and looks for the actual search term in these sequences, allowing phoneme insertions, deletions and substitutions. These operations have different costs depending on the given phoneme (or phoneme pair), calculated from phoneme-level confusion statistics.

We used recordings of Hungarian broadcast news for testing, which were taken from 8 different TV channels [5]. The 70 broadcast news recordings were divided into three groups: the first, largest one (about 5 hours long) was used for training purposes. The second part (about an hour long) was the *development set*: these recordings were used to determine the optimal threshold for the ATWV metric. The third part was the *test set* (about 2 hours long), and it was used to evaluate the overall performance. We chose 50 words and expressions as search terms, which came up in the news recordings quite frequently. They varied between 6-16 phonemes in length (2-6 syllables), and they were all nouns, one-third of them (18) being proper nouns. The phoneme sequence intermediate representations were produced by Artificial Neural Networks [2], trained in the way described in [18], using the standard MFCC $+\Delta + \Delta\Delta$ feature set [8].

3.2 Results

The accuracy scores produced by our actual STD system (using different strategies for determining the list of relevant occurrences) can be seen in Table 2. By "Automatic" we mean the standard, automatic method used for determining correct hits; "Subject #*N*" means the responses of the *N*th subject. Below we list the mean and the median values of the accuracy scores produced, and the scores obtained using majority voting. The last line shows the accuracy scores calculated without any subject answers, using just the clean occurrences; that is, in this case we treated an occurrence as a correct one only if the keyword appeared unchanged in the transcription at the beginning of a word.

The first thing to notice is that the FOM scores practically do not vary, which is probably due to the way this accuracy score is calculated: it is relatively easy to achieve high FOM scores, but it is very hard to significantly improve them. The ATWV scores, however, differ much more from each other, ranging from 48.00% (where we use only the clean occurrences) to 60.23% when using the list of relevant occurrences given by Subject #3. The results are also quite different from the case where we applied our automatic method.

Strategy	FOM	ATWV	F_1	Prec.	Recall
Automatic	88.72%	56.84%	85.29%	91.17%	80.11%
Subject #1	88.35%	52.32%	83.93%	88.44%	79.86%
Subject #2	87.39%	48.00%	82.32%	86.68%	78.37%
Subject #3	88.85%	60.23%	86.05%	93.58%	79.64%
Subject #4	88.15%	52.90%	84.11%	89.25%	79.54%
Subject #5	88.22%	53.05%	84.24%	89.25%	79.77%
Subjects (mean)	88.19%	53.30%	84.13%	89.44%	79.44%
Subjects (median)	88.22%	52.90%	84.11%	89.25%	79.64%
Subjects (majority voting)	88.22%	53.07%	84.24%	89.25%	79.77%
Clean occurrences	87.94%	44.77%	81.48%	83.31%	79.72%

Table 2. STD accuracy scores using different strategies for determining correct hits

The F_1 scores varied from 82.32% to 86.05%. Quite interestingly, the corresponding precision scores were practically the same, so the difference came from the recall scores. The correlation of the precision, F-measure, ATWV scores, and the number of occurrences marked as real is clear: for Subject #3 these were 93.58%, 86.05%, 60.23% and 732, respectively, whereas for Subject #2 these were 86.68%, 82.32%, 48.00% and 689. (The ATWV metric is known to be fairly sensitive to false alarms.)

Another interesting finding is that the scores belonging to majority voting appear to be quite close to those of three subjects (#1, #4 and #5), or the mean/median of all the subjects. This suggests that by using the simple technique of majority voting a consensus of correct hits can be achieved, which falls quite close to the expectations of the average user.

3.3 Verifying the Occurrences

Having evaluated the accuracy scores belonging to the different subject responses, we will now turn to the perhaps more interesting part, where we focus on the more significant and/or more interesting differences among the responses of the users or between the user-entered and the automatic hit lists. Note that, as we used a Hungarian database for this study, the examples below will also be in Hungarian; nevertheless, we think that the cases encountered have a much wider scope as probably quite similar types appear in other languages as well.

One well-known drawback of language-independent STD approaches is that they are likely to produce false alarms when the (usually short) actual search term is contained inside another word. In our case, one such example was the term *"kormány"* (meaning *cabinet*), which came up quite frequently inside the word *"önkormányzat"* (*local council*). Since in this case the whole keyword is present, the automatic occurrence detector method included these as real occurrences, whereas 4 of the 5 subjects treated them as false alarms. Of course the STD system, relying only on the acoustic data, also found these occurrences.

Recall that, due to the agglutinative property of the Hungarian language, we allowed the final vowel of the keyword (as long as it was also the last phoneme) to change to its long counterpart, so the STD system was also expected to find these occurrences. However, by default no such changes with earlier vowels were allowed, although they were also sometimes related to similar word-pairs. A good example of this is the keyword *"vasút"* (meaning *railway*) and the word *"vasutas"* (*railway worker*); each subject viewed the latter word as a relevant occurrence of the search term. Yet, for the term *"miniszter"* (*minister*), there is only a vowel difference in *"minisztérium"* (*ministry*), hence it is exactly the same type as the previous one; but it was rejected by 4 out of the 5 subjects.

Another big group was the presence of certain proper nouns in the list of keywords, typically names of people like *"Angela Merkel"* (German chancellor), *"Bajnai Gordon"* or *"Orbán Viktor"* (both of them being Hungarian prime ministers[1]). The search terms consisted of their full names (i.e. both first and family names), whereas sometimes these people were referred to only by their family names. All the subjects agreed that these were real occurrences, despite that only half of the actual keywords were present at the given position. Note that as we used edit distance when creating the form, only those occurrences were present for the subjects to evaluate where the context was sufficiently similar to the first name (e.g. *"amely Merkel"*, "*Bajnai kormány"*, *"Orbán kormány"*).

A quite similar case was that of the keyword *"rendőrség"* (*police force*), which, due to the similarity of the word following it, proved likely to occur in a recording where only the word *"rendőr"* (*policeman*) was present. Here 3 of the 5 subjects found this "inverse containment" relevant, indicating that the concept of the two words are strongly related. In the last frequent case the keyword was *"gázár"* (*gas price*), and the listed items in the form all contained *"gáz ára"* (*price of gas*); all subjects thought that these were real occurrences of the search term.

From these examples it can be seen that the subjects usually agreed with each other, but their choice can hardly be predicted automatically. If a word contains the keyword, then it is usually a correct occurrence. But at certain times (*kormány*) it is a false alarm, while at other times (*rendőrség*) the keyword contains the word that actually occurred. The last vowel of the keyword may become its long counterpart. But such a change is sometimes allowed for other vowels as well (*vasút*), while sometimes it is not (*miniszter*). The case of *"gázár"* probably cannot be handled at all: allowing word boundaries inside keywords would lead to a lot of false alarms. Still, when looking for famous people, the keyword should be only their family name (like *Merkel*, *Bajnai* and *Orbán*).

The accuracy scores in Table 2 also accord with our findings when examining the actual answers of subjects. Subject #3 accepted both *"minisztérium"* for the keyword *"miniszter"* and *"önkormányzat"* for the search term *"kormány"*; this compliance reduced the number of false alarms for the STD system, leading to high precision, ATWV and F_1 scores. In contrast, Subject #2 rejected several compound words as correct hits, which were all accepted by the other four subjects; this is also reflected in the lower precision, F_1 and ATWV scores.

[1] Although, of course, not at the same time

Quite interestingly, when there was a disagreement among the subjects, in most cases four of them agreed on one option, and only in four instances was there a voting outcome of three to two. This may indicate that in almost every case a broad consensus can be achieved, although this should be tested in experiments with more subjects. Our test results also support this hypothesis: increasing the number of votes required to four lowered the accuracy scores only slightly, whereas when we required that all subjects should agree, they fell more sharply.

Comparing the scores obtained involving human interaction with those we got using the two automatic methods to determine the relevant occurrences, it is clear that they differ significantly: when we only allowed the clean hits, the resulting ATWV score of 44.77% was low compared to the others due to the high number of false alarms; whereas when we used the standard automatic method, it was too permissive, resulting in an overoptimistic ATWV score of 56.84%.

Based on these observations, we can sum up our findings in three parts. Firstly, keyword selection should match user behaviour a bit more: all search terms should be nouns, preferably proper nouns (e.g. names, cities, etc.), and for well-known people only their family name should be used. Of course a limitation for this is the set of available recordings (so that the given keywords should occur in the dataset several times); still, further investigations should be preceded by a more careful keyword selection.

The form containing the possible occurrences was constructed in a syntactical manner (using the edit distance-based similarity of the transcriptions); from the results it seems that we should also turn to a linguistic analysis. It would mean a more robust way to distinguish, for example, the inflected forms (e.g. plurals) of the keywords from compound words, since the latter ones should remain in the form to fill, whereas the former occurrences should be omitted.

Overall, it seems that the users focus on the stem of the keywords, often even dismissing affixes (e.g. *rendőr* instead of *rendőrség*, *vasút* instead of *vasutas*). In some cases this is also an oversimplification (e.g. the case of *miniszter* – *minisztérium*), but it still seems to be a pretty close estimation of keyword occurrence relevance. A deeper analysis could be performed via a more detailed linguistic analysis like using Natural Language Processing tools, or expressing the type of connection between word forms via a WordNet [11, 10].

4 Conclusions

In this study, we examined the spoken term detection task from an unusual viewpoint: we checked how much automatically generated ground truth keyword occurrences match user expectations. For this, we asked a number of subjects to mark the possible occurrences that they thought were relevant. We found that although no two subjects gave exactly the same responses, generally their answers were quite similar; and by using majority voting a clear consensus could be achieved. But the standard automatic keyword occurrence detection methods used were either too lax or too strict when compared with the subject responses.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York (1999)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
3. Blair, D.C., Maron, M.E.: An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM 28(3), 289–299 (1985)
4. Farkas, R., Vincze, V., Nagy, I., Ormándi, R., Szarvas, G., Almási, A.: Web-based lemmatisation of named entities. In: Proceedings of TSD. pp. 53–60. Brno, Czech Republic (2008)
5. Gosztolya, G., Tóth, L.: Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal (in Hungarian). In: Proceedings of MSZNY. pp. 224–235. Szeged, Hungary (2010)
6. Gosztolya, G., Tóth, L.: Spoken term detection based on the most probable phoneme sequence. In: Proceedings of SAMI. pp. 101–106. Smolenice, Slovakia (Jan 2011)
7. Hakkani-Tür, D.Z., Oflazer, K., Tür, G.: Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities 36(4), 381–410 (2002)
8. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice Hall (2001)
9. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
10. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of GWC. pp. 310–320 (2008)
11. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
12. NIST: Spoken Term Detection 2006 Evaluation Plan (2006)
13. Özgür, A., Özgür, L., Güngör, T.: Text categorization with class-based and corpus-based keyword selection. In: Proceedings of ISCIS. pp. 607–616 (2005)
14. Parlak, S., Saraclar, M.: Spoken term detection for Turkish broadcast news. In: Proceedings of ICASSP. pp. 5244–5247 (2008)
15. Pinto, J., Hermansky, H., Szöke, I., Prasanna, S.: Fast approximate spoken term detection from sequence of phonemes. In: Proceedings of SIGIR. Singapore (2008)
16. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall (1993)
17. Tangruamsub, S., Punyabukkana, P., Suchato, A.: Thai speech keyword spotting using heterogeneous acoustic modeling. In: Proceedings of RIVF. pp. 253–260 (2007)
18. Tóth, L.: A hierarchical, context-dependent Neural Network architecture for improved phone recognition. In: Proceedings of ICASSP. pp. 5040–5043 (2011)
19. Wang, D.: Out-of-Vocabulary Spoken Term Detection. Ph.D. thesis, University of Edinburgh (2010)
20. Xu, J.W., Singh, V., Govindaraju, V., Neogi, D.: A hierarchical classification model for document categorization. In: Proceedings of ICDAR. pp. 486–490 (2009)
21. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)
22. Zhang, P., Han, J., Shao, J., Yan, Y.: A new keyword spotting approach for spontaneous Mandarin speech. In: Proceedings of ICSP (2006)